Multi-Label Learning with Multiple Complementary Labels

Yi Gao, Jing-Yi Zhu, Miao Xu, and Min-Ling Zhang, *Senior Member, IEEE*

Abstract—In *multi-labeled complementary label learning* (MLCLL), a *complementary label* (CL) represents an irrelevant label for an instance. Utilizing CLs instead of relevant labels as annotations simplifies the annotation process in *multi-label learning* (MLL) tasks, underscoring the practicality of the MLCLL problem. However, existing MLCLL approaches mainly focus on scenarios where an instance is associated with a single CL. This restricts their applicability in situations where annotators provide multiple CLs per instance. To address this limitation, we propose a novel paradigm called multi-label learning with multiple complementary labels (ML-MCL), which allows each instance to be associated with multiple CLs simultaneously. Through analyzing the generation process of multiple CLs, we construct the relationship between relevant labels and CLs. This assists in deriving a tailored risk-consistent estimator to solve MLCLL with multiple CLs. Theoretically, we establish an estimation error bound for this estimator, with a convergence rate of $\mathcal{O}(1/\sqrt{n})$. Furthermore, we observed that unbounded gradients can be produced in the derived estimator when optimizing with certain loss functions, which may lead to unstable optimization. To mitigate this issue, we enhance the estimator with a *confidence truncation* loss, stabilizing the optimization process. Experimental results confirm the effectiveness of our approach, showing improved learning stability and performance in MLCLL tasks involving multiple CLs.

Index Terms—Complementary label learning, multi-label learning, risk-consistent estimator, estimation error bound.

1 Introduction

Multi-label learning (MLL) aims to learn a multi-labeled classifier that can assign multiple relevant labels to an unseen instance simultaneously [1], [2], [3]. However, fully supervised MLL tasks generally require massive precisely multi-labeled data, the collection of which is expensive and laborious [4], [5], [6]. To alleviate this problem, many researchers turn to study weakly supervised learning, which enables learning under weak supervision information [7]. At present, various weakly supervised learning frameworks have been widely studied, including but not limited to, *one positive label learning* [8], [9], *semi-supervised MLL* [10], [11], *positive-unlabeled learning for MLL* [12], [13], and *partial multi-label learning* (PML) [14], [15].

Here, we explore another weakly supervised learning scenario termed *multi-labeled complementary label learning* (MLCLL) [4], [16]. In MLCLL, each training instance is associated with a single *complementary label* (CL), which specifies an irrelevant label for that instance. The goal of MLCLL is identical to that of MLL, which is to learn a multi-labeled classifier capable of assigning a set of relevant labels to an unseen instance. Obviously, collecting CLs is less laborious than collecting multiple precise relevant labels. This simplifies the annotation process by circumventing complex semantic labels and the unknown number of relevant labels in fully supervised MLL [4], [16]. Moreover, the collection of CLs avoids the need for annotators to check all relevant labels for instances individually across the entire label space. However, the existing MLCLL problem allows only a single CL for each instance, which significantly restricts its potential. This constraint provides an opportunity to expand this paradigm, as in real-world scenarios, annotators may often provide multiple CLs for an instance, enhancing the richness of the training data.

Recently, CLs have been applied in the medical domain [17], [18]. For example, healthcare professionals frequently encounter patients exhibiting symptoms that could indicate multiple health conditions. The multi-label nature of medical diagnoses necessitates the identification of several potential diseases simultaneously, which is inherently challenging. CLs prove valuable in such scenarios by allowing medical experts to confidently exclude certain diseases based on observed symptoms. By systematically eliminating less likely conditions (i.e., irrelevant labels serving as CLs), physicians can more accurately infer the most probable diagnoses, thereby enhancing the diagnostic process. Beyond healthcare, CLs are also useful in e-commerce product categorization, where annotators may find it easier to exclude irrelevant categories rather than assign all appropriate ones. In this case, multiple CLs would be more commonly provided than a single CL [19].

In this paper, we propose a novel paradigm, called multi-label learning with multiple complementary labels (ML-MCL), which enables each instance to be associated with multiple CLs simultaneously. While existing MLCLL approaches, as discussed in previous studies [4], [16], have shown promising results on a single CL for an instance, their efficacy in scenarios involving multiple CLs is yet to be established. For instance, Gao et al. [16] recovered the true multi-labeled data distribution from complementary

- *Yi Gao and Min-Ling Zhang (corresponding author) are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. E-mail: gao_yi@seu.edu.cn; zhangml@seu.edu.cn*
- *Jing-Yi Zhu is with the Suzhou Joint Graduate School, Southeast University, China. E-mail: zhujingyi@seu.edu.cn*
- *Miao Xu is with the University of Queensland, Australia. E-mail: miao.xu@uq.edu.au*

labeled data by assuming that each instance is related to only a single CL; in [4], a transition matrix for MLCLL was estimated under the assumption that only a single CL is available for each instance. These approaches are thus inherently limited, as they rely on the premise of a single CL per instance, which restricts their applicability in learning scenarios where multiple CLs are present.

To address the problem of MLCLL with multiple CLs, we begin by analyzing the generation process of multiple CLs for ML-MCL. This analysis not only clarifies the data distribution but also deepens our understanding of how complementary labeled instances are generated. This process allows us to establish the relationship between relevant labels and their corresponding CLs, which allows us to derive a risk-consistent estimator. This estimator guarantees that the classifier learned from multiple CLs converges to the optimal one achievable under fully supervised MLL. Theoretically, we establish an estimation error bound for our proposed risk estimator and prove its convergence rate. Furthermore, we observe that the risk-consistent estimator with certain loss functions may result in unbounded gradients, which causes instability in the training process. To alleviate this issue, we enhance the risk estimator by minimizing a *confidence truncated loss* (CTL) designed specifically in this paper. This improvement not only benefits gradient updates but also stabilizes the optimization process. Our experimental results demonstrate the effectiveness of this approach. The main contributions of our work can be summarized as follows:

- We propose a novel paradigm called ML-MCL, which allows learning with multiple CLs. To address this new paradigm, we derive a risk-consistent estimator by analyzing the generation process of multiple CLs.
- The risk-consistent estimator ensures that the classifier learned from multiple CLs will converge to the optimal one in fully supervised MLL. We establish an estimation error bound for the proposed risk estimator, with a convergence rate of $\mathcal{O}(1/\sqrt{n})$.
- To solve issues related to the unstable learning process caused by unbounded gradients, we design CTL to improve our risk estimator. This improvement further facilitates gradient updates and stabilizes the optimization process.

The remaining organization of this paper is as follows. Section 2 provides a brief review of related work. Sections 3 and 4 describe the proposed approach and CTL, respectively. Section 5 presents the experimental results, and Section 6 concludes the paper.

2 Related Work

In this section, we briefly review related work on ML-MCL, including MLL, complementary label learning in multi-class classification, and MLCLL.

2.1 Multi-Label Learning

In fully supervised MLL, each instance is equipped with a set of relevant labels. The goal is to learn a multi-label classifier that can assign relevant labels to unseen instances. Existing MLL approaches can be categorized into three research lines based on the order of label correlations: first-order approaches [20], second-order approaches [21], [22], and high-order approaches [23]. First-order approaches solve MLL problems by decomposing them into a series of binary classification tasks [20], which disregard the relationship among labels. However, researchers have found that label correlations exist in multi-labeled data [1], [2]. Therefore, many studies have turned to consider label correlations to solve MLL problems. Second-order approaches focus on the correlations between pairs of labels [21], [22]. These methods typically convert MLL problems into bipartite ranking problems by ensuring that relevant labels are ranked higher than irrelevant ones [24], [25]. On the other hand, high-order approaches explore more complex relationships than second-order ones, exploiting label correlations among label subset or all labels in the label space [26], [27], [28]. Although high-order approaches can model stronger label correlations, they incur higher computational costs compared to first and second-order approaches [29]. In addition, ML-MCL is more challenging than MLL tasks because it lacks access to relevant labels. As a result, conventional MLL approaches struggle to handle this paradigm effectively.

2.2 Complementary Label Learning in Multi-Class Classification

In multi-class classification, each instance is equipped with a relevant label. Similar to MLL problems, collecting high-quality labeled data is hard in multi-class learning [30]. To address this issue, complementary label learning was first proposed as a solution to the difficulty of obtaining precisely labeled data in multi-class learning [31]. In their pioneering work, Ishida et al. [31] derived an unbiased risk estimator under a uniform assumption and reformulated one-versus-all and pairwise comparison loss functions to address the problem. To mitigate the limitation of being constrained to specific loss functions, Ishida et al. [32] proposed a new framework that can accommodate arbitrary loss functions and models. For enhanced practicality, biased CLs have been explored by estimating a transition matrix [33], [34]. However, approaches based on transition matrices require additional conditions, such as the availability of anchor instances, limiting their suitability for real-world scenarios. To ease dependence on an estimated transition matrix, Gao et al. [35] directly modeled the probabilities of CLs using the model's outputs. The effectiveness of these approaches relies on the assumption that each instance has only a single CL. Consequently, they may encounter challenges in solving the problem setting of ML-MCL, as the number of relevant labels per instance in MLL is unknown and can vary across instances.

2.3 Multi-Labeled Complementary Label Learning

To alleviate the challenges of collecting precisely multi-labeled data in MLL, the problem of MLCLL has attracted many researchers to investigate. Existing MLCLL approaches mainly focus on scenarios where each instance is equipped with only a single CL [4], [16]. For example, Gao et al. [16] recovered relevant labels from complementary labeled data based on a uniform generation assumption, where each instance is annotated with a single CL. They also designed a *gradient-descent friendly* (GDF) loss function to boost the model's performance. In another study [4], a transition matrix-based approach was proposed, which reconstructed relevant labels by estimating a transition matrix under the assumption that each instance is associated with a single CL. However, in real-world scenarios, annotators may often provide multiple CLs for an instance. This limitation, where each instance is assumed to have only a single CL, heavily restricts the applicability of existing MLCLL approaches in cases with multiple CLs. Existing MLCLL approaches primarily focus on single CL scenarios and do not fully address the complex dynamics of multiple CLs. Therefore, to overcome this limitation, we propose a risk-consistent estimator capable of ML-MCL.

3 Learning with Multiple Complementary Labels

In this section, we first introduce the notations and problem setting. By analyzing the generation process of multiple CLs, we establish the relationship between relevant labels and CLs. This relationship allows us to recover the relevant label distribution from complementary labeled data and to derive a risk-consistent estimator, along with its estimation error bound.

3.1 Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be the feature space with d dimensions, and $\mathcal{Y} = \{1, 2, 3, \ldots, K\}$ be the label space with K possible labels, where $K > 2$. In fully supervised MLL, we define $\boldsymbol{x}$ as an instance, and Y as the set of relevant labels for the instance $\boldsymbol{x}$. Here, we assume that $(\boldsymbol{x}, Y) \in (\mathcal{X}, \mathcal{Y})$ is independently drawn from an unknown joint probability distribution $p(\boldsymbol{x}, Y)$. The goal of MLL is to learn a classifier $\boldsymbol{f} : \mathcal{X} \mapsto [0,1]^K$ that can assign predictions for unseen instances. The classifier $\boldsymbol{f}$ is obtained by minimizing the following expected classification risk:

$$R(\boldsymbol{f}) = \mathbb{E}_{p(\boldsymbol{x},Y)}[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)], \tag{1}$$

where $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)$ refers to MLL loss functions, defined as $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y) = \sum_{y=1, y\in\mathcal{Y}}^{K} \ell_y(\boldsymbol{x}) + \sum_{y=1, y\notin Y}^{K} \bar{\ell}_y(\boldsymbol{x})$. We define $f_y(\cdot)$ as the y-th prediction of $\boldsymbol{f}(\cdot)$, which is used to estimate $p(y = 1|\boldsymbol{x})$. $\ell_y(\boldsymbol{x})$ and $\bar{\ell}_y(\boldsymbol{x})$ calculate the loss of $f_y(\boldsymbol{x})$ when y belongs to the relevant and irrelevant labels, respectively. Specifically, when $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)$ refers to *Binary Cross Entropy* (BCE) loss, $\ell_y(\boldsymbol{x}) = -\log(f_y(\boldsymbol{x}))$ and $\bar{\ell}_y(\boldsymbol{x}) = -\log(1 - f_y(\boldsymbol{x}))$. Additionally, $\ell_y(\boldsymbol{x}) = 1 - f_y(\boldsymbol{x})$ and $\bar{\ell}_y(\boldsymbol{x}) = f_y(\boldsymbol{x})$ when *mean absolute error* (MAE) loss is used by $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)$. Note that an approach is risk-consistent if its learned classification risk estimator equals $R(\boldsymbol{f})$ given the same classifier $\boldsymbol{f}$ [8], [36], [37].

In this paper, we study the problem of learning with multiple CLs in MLCLL, namely ML-MCL. Given a complementary labeled dataset $\bar{D} = \{(\boldsymbol{x}_i, \bar{Y}_i)\}_{i=1}^{n}$ consisting of n instances, each independently sampled from an unknown joint probability distribution $\bar{p}(\boldsymbol{x}, \bar{Y})$, where $\bar{Y}_i \subseteq \mathcal{Y} - Y_i$ represents a set of CLs for an instance $\boldsymbol{x}_i \in \mathcal{X}$. Specifically, $\bar{Y}_i$ is a subset chosen from the remaining labels after removing the relevant label set Y_i from the label space $\mathcal{Y}$. It is important to note that $\bar{Y}_i$ cannot be an empty set nor the full label set, which ensures the validity of our problem setting. Thus, $\bar{Y} \in \bar{\mathcal{Y}}$, where $\bar{\mathcal{Y}} = \{2^{\mathcal{Y}} - \emptyset - \mathcal{Y}\}$. The goal of ML-MCL is the same as in MLL, which is to learn a multi-labeled classifier $\boldsymbol{f} : \mathcal{X} \mapsto [0,1]^K$. This task extends the existing MLCLL framework from the scenario of a single CL to multiple CLs, thereby enhancing its applicability to real-world situations. However, ML-MCL introduces additional challenges due to the uncertain and variable number of CLs across instances. Due to the complex dynamics of multiple CLs, dealing with multiple CLs makes this framework more challenging than the single CL case. In the next subsection, we will analyze the generation process of multiple CLs to help construct the relationship between multiple CLs and relevant labels, which will further aid in deriving a risk-consistent estimator.

3.2 Data Generation Process

Without any additional knowledge, inferring the generation process of multiple CLs is difficult as the number of relevant labels and CLs is uncertain. Motivated by Feng et al. [19], we assume that the generation process relies on the size of the set of multiple CLs. Let's denote the size of the complementary label set as a random variable s, where s follows a distribution $p(s)$. Under this umbrella, we assume each training instance $(\boldsymbol{x}_i, \bar{Y}_i)$ is drawn from $\bar{p}(\boldsymbol{x}, \bar{Y})$, which is defined as:

$$\bar{p}(\boldsymbol{x}, \bar{Y}) = \sum_{j=1}^{K-1} p(s = j)\bar{p}(\boldsymbol{x}, \bar{Y}|s = j), \tag{2}$$

where $s \neq 0, K$ and $\bar{p}(\boldsymbol{x}, \bar{Y}|s = j) :=$

$$\begin{cases} \sum\limits_{Y\in\mathcal{Y}, Y\cap\bar{Y}=\emptyset} \frac{1}{C_{K-|Y|}^{j}} p(\boldsymbol{x}, Y), & \text{if } |\bar{Y}| = j, j \leq K - |Y| \\ 0, & \text{otherwise} \end{cases}.$$

Obviously, this distribution will simplify to the MLCLL problem with a single CL when $p(s = 1) = 1$ [16]. Eq. (2) specifies the probability of each set of multiple CLs being uniformly sampled, given Y. Additionally, Eq. (2) reveals the relationship between relevant labels and multiple CLs with certain constraints. Specifically, the selection of $\bar{Y}_i$ is influenced by Y_i, ensuring that $Y_i \cap \bar{Y}_i = \emptyset$. This modeling choice explicitly prevents overlap between $\bar{Y}_i$ and Y_i. In Theorem 1, we will show the validity of our assumed probability distribution $\bar{p}(\boldsymbol{x}, \bar{Y})$ by establishing the necessary conditions that ensure $\bar{p}(\boldsymbol{x}, \bar{Y})$ constitutes a valid probability distribution. This validation further guarantees that the overall probability model is well-defined.

Theorem 1. *$\bar{p}(\boldsymbol{x}, \bar{Y})$ is a valid probability distribution, which satisfies non-negativity and $\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{Y})}[1] = 1$.*

The proof is provided in Appendix A. Here, we present a real-world motivation for the assumed data distribution. In real-world data collection, we can approximate Eq. (2) by first randomly sampling the size s from $p(s)$ and then uniformly selecting s labels from the entire label space to form a candidate set. If annotators confirm that none of these labels are correct for the given instance, this set is treated as a set of multiple CLs, aligning with the distribution described in Eq. (2). For example, in medical diagnostics, a physician might be presented with a small subset of potential diagnoses. If the physician determines that none of these apply to the patient, the subset serves as a complementary label set. This approach not only approximates the assumptions but also alleviates the annotation burden, as it is generally easier for experts to exclude incorrect options rather than identify all relevant ones.

Furthermore, Eq. (2) describes the generation process of multiple CLs from $p(\boldsymbol{x}, Y)$, while directly recovering relevant labels from multiple CLs based on Eq. (2) is not feasible since the multi-labeled data is unavailable at here. Hence, we proceed to derive a risk-consistent estimator by investigating Lemma 2.

Lemma 2. *Let $\bar{\mathcal{Y}}_j = \{\bar{Y} | \bar{Y} \in \bar{\mathcal{Y}}, |\bar{Y}| = j\}$. With Eq. (2),*

$$p(\boldsymbol{x}, Y) = \frac{1}{2^K - 2} \sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \bar{p}(\boldsymbol{x}, \bar{Y} | s = j). \tag{3}$$

The proof is stated in Appendix B. This lemma constructs the MLL probability distribution from the complementary labeled data.

3.3 The Risk-Consistent Estimator

A risk-consistent estimator allows the evaluation of the classification risk of fully supervised MLL using data that is only associated with multiple CLs. Based on Lemma 2, the following theorem shows a risk-consistent estimator that is equivalent to Eq. (1) when given the same classifier.

Theorem 3. *Under Lemma 2, $R(\boldsymbol{f}) = \bar{R}(\boldsymbol{f})$ based on the definitions of $\bar{p}(\boldsymbol{x}, \bar{Y})$ and $R(\boldsymbol{f})$. $\bar{R}(\boldsymbol{f})$ is expressed as:*

$$\bar{R}(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s = j) \bar{R}_j(\boldsymbol{f}), \tag{4}$$

where $\bar{R}_j(\boldsymbol{f}) = \mathbb{E}_{\bar{p}(\boldsymbol{x}, \bar{Y} | s=j)} \left[\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}), \bar{Y}) \right]$ and

$$\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}), \bar{Y}) = \frac{2^{K-j-1}}{2^K - 2} \sum_{y=1, y \notin \bar{Y}}^{K} \ell_y(\boldsymbol{x}) + \frac{2^{K-j-1} - 1}{2^K - 2} \sum_{y=1, y \notin \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x}) + \frac{2^{K-j} - 1}{2^K - 2} \sum_{y=1, y \in \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x}).$$

The proof is provided in Appendix C. Theorem 3 shows that the fully supervised classification risk can be estimated by the risk-consistent estimator $\bar{R}(\boldsymbol{f})$ with the corresponding loss using complementary data. Moreover, it ensures that the learned classifier in $\bar{R}(\boldsymbol{f})$ will converge to $R(\boldsymbol{f})$. Since the probability distribution $\bar{p}(\boldsymbol{x}, \bar{Y})$ is generally unknown even with the complementary labeled dataset, the expected risk $\bar{R}(\boldsymbol{f})$ is usually approximated by the empirical risk $\bar{R}_n(\boldsymbol{f})$, i.e.,

$$\bar{R}_n(\boldsymbol{f}) = \sum_{j=1}^{K-1} \frac{p(s = j)}{n_j} \sum_{i=1}^{n_j} \bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}_i), \bar{Y}_i) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}_i), \bar{Y}_i), \tag{5}$$

where $\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}_i), \bar{Y}_i) = \Big\{ 2^{K-|\bar{Y}_i|-1} \sum_{y=1, y \notin \bar{Y}}^{K} \ell_y(\boldsymbol{x}_i) + (2^{K-|\bar{Y}_i|-1} - 1) \sum_{y=1, y \notin \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x}_i) + (2^{K-|\bar{Y}_i|} - 1) \sum_{y=1, y \in \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x}_i) \Big\} / (2^K - 2)$. We can empirically approximate $p(s = j)$ by n_j / n, where n_j refers to the number of instances in $\bar{D}$ whose size of CLs is j.

3.4 Estimation Error Bound

We establish an estimation error bound for our proposed approach based on Rademacher Complexity [38] to verify the convergence rate. Let $\mathcal{F}$ be the hypothesis class, and $\mathcal{G}_y = \{g : \boldsymbol{x} \mapsto f_y(\boldsymbol{x}) | \boldsymbol{f} \in \mathcal{F}\}$ be the functional space for the label $y \in \mathcal{Y}$. $\mathfrak{R}_n(\mathcal{G}_y)$ indicates Rademacher Complexity of $\mathcal{G}_y$, which is defined as $\mathfrak{R}_n(\mathcal{G}_y) = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}_y} \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) \right]$. Assuming $\boldsymbol{f}_n = \arg\min_{\boldsymbol{f} \in \mathcal{F}} \bar{R}_n(\boldsymbol{f})$ is the empirical risk minimizer, and $\boldsymbol{f}^* = \arg\min_{\boldsymbol{f} \in \mathcal{F}} R(\boldsymbol{f})$ is the true risk minimizer, we present the following theorem.

Theorem 4. *Let $M_j = \sup_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{f} \in \mathcal{F}} \bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}), \bar{Y})$. For any $y \in \mathcal{Y}$, assuming $\ell_y(\boldsymbol{x})$ and $\bar{\ell}_y(\boldsymbol{x})$ are ρ^+-Lipschitz and ρ^--Lipschitz with respective to $\boldsymbol{f}(\boldsymbol{x})$, respectively. For any $\delta > 0$, with a probability at least $1 - \delta$,*

$$R(\boldsymbol{f}_n) - R(\boldsymbol{f}^*) \leq \sum_{j=1}^{K-1} p(s = j) \left\{ 4\sqrt{2} K \mathcal{C}_j \sum_{y=1}^{K} \mathfrak{R}_{n_j}(\mathcal{G}_y) + M_j \sqrt{\frac{\log 2/\delta}{2 n_j}} \right\}, \tag{6}$$

where $\mathcal{C}_j = \frac{2^{K-j-1}}{2^K - 2} \rho^+ + \frac{3 \cdot 2^{K-j-1} - 2}{2^K - 2} \rho^-$ for all $j \in \{1, 2, \ldots, K-1\}$.

The proof is provided in Appendix D. Theorem 4 demonstrates that the proposed risk-consistent estimator possesses an estimation error bound with a convergence rate of $\mathcal{O}(1/\sqrt{n})$. It further indicates that the gap between $R(\boldsymbol{f}_n)$ and $R(\boldsymbol{f}^*)$ is close to 0 as $n \to \infty$, which signifies the convergence of the empirical minimizer to the true risk minimizer. Notably, the distribution shown in Eq. (2) will be simplified to [16] when the number of CLs per instance is $|\bar{Y}| = 1$. Consequently, our estimator naturally reduces to the form proposed in [16], and our estimation error bound aligns with that of [16], indicating that the single-CL setting is a special case of our generalized framework. While our data generation assumptions are inspired by [19], since [19] does not explicitly provide a general framework for MLL, its estimator and estimation error bound cannot be directly extended to our setting in the same manner as [16].

4 Confidence Truncated Loss

As discussed earlier, we derive a risk-consistent estimator that can accommodate arbitrary loss functions in MLL according to the generation process of multiple CLs. This naturally raises two questions:

1) What impact will different loss functions have on the risk-consistent estimator?
2) Which type of loss function is most beneficial for optimizing our risk-consistent estimator?

In this section, we proceed by examining several MLL loss functions, including the BCE loss and MAE loss, to investigate their impacts on our risk estimator from the perspective of gradients. Building on this analysis, we will explore the loss functions that enhance the optimization of the risk-consistent estimator.

Next, we explore the situation of the BCE loss to the risk-consistent estimator by introducing the BCE loss into $\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}), \bar{Y})$. We denote this as $\bar{\mathcal{L}}_{\mathrm{BCE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{Y})$. which is defined as:

$$\begin{aligned}\bar{\mathcal{L}}_{\mathrm{BCE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{Y}) = &-\frac{2^{K-|\bar{Y}|-1}}{2^K-2}\sum_{y=1, y\notin\bar{Y}}^{K}\log(f_y(\boldsymbol{x})) \\ &-\frac{2^{K-|\bar{Y}|-1}-1}{2^K-2}\sum_{y=1, y\notin\bar{Y}}^{K}\log(1-f_y(\boldsymbol{x})) \\ &-\frac{2^{K-|\bar{Y}|}-1}{2^K-2}\sum_{y=1, y\in\bar{Y}}^{K}\log(1-f_y(\boldsymbol{x})).\end{aligned} \tag{7}$$

The gradient of $\bar{\mathcal{L}}_{\mathrm{BCE}}$ with respect to $\boldsymbol{\theta}$, the learnable parameters for $f_y(\boldsymbol{x})$, is given by:

$$\frac{\partial\bar{\mathcal{L}}_{\mathrm{BCE}}}{\partial\boldsymbol{\theta}} = \begin{cases}(w^- - w^+)\,\nabla_{\boldsymbol{\theta}}\, f_y(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } y\notin\bar{Y} \\ \frac{2^{K-|\bar{Y}|}-1}{2^K-2}\,\frac{\nabla_{\boldsymbol{\theta}} f_y(\boldsymbol{x};\boldsymbol{\theta})}{1-f_y(\boldsymbol{x};\boldsymbol{\theta})}, & \text{if } y\in\bar{Y}\end{cases}, \tag{8}$$

where $w^+ = 2^{K-|\bar{Y}|-1}/[(2^K-2)f_y(\boldsymbol{x};\boldsymbol{\theta})]$ and $w^- = (2^{K-|\bar{Y}|-1}-1)/[(2^K-2)(1-f_y(\boldsymbol{x};\boldsymbol{\theta}))]$. The calculation of Eq. (8) can be divided into two parts: CLs ($\bar{Y}$) and non-CLs ($\mathcal{Y}-\bar{Y}$). For CLs, gradient descent benefits when the prediction $f_y(\boldsymbol{x};\boldsymbol{\theta})$ is close to zero, especially when the label y belongs exclusively to $\bar{Y}$. However, for non-CLs, the situation can lead to infinite values for w^- or w^+, regardless of how close the label prediction is to the groundtruth (0 or 1). For example, imagine there is another label y_o that is irrelevant to $\boldsymbol{x}$ and belong to non-CLs, such that its prediction $f_{y_o}(\boldsymbol{x};\boldsymbol{\theta})$ will be close to zero. If $f_{y_o}(\boldsymbol{x};\boldsymbol{\theta}) = 0$ or close to zero, w^+ in Eq. (8) becomes infinite, leading to an unbounded gradient even if f_{y_o} is close to the groundtruth. Conversely, if a label y_r, a relevant label of $\boldsymbol{x}$, has a groundtruth prediction close to one, it will result in an infinite w^- and thus an infinite gradient. These examples illustrate that infinite values of w^+ or w^- can cause unbounded gradients when using the risk-consistent estimator, potentially resulting in unstable optimization.

Similarly, we start by exploring the impact of MAE loss on the risk estimator through analyzing the gradient of $\bar{\mathcal{L}}_{\mathrm{MAE}}$ with respect to $\boldsymbol{\theta}$. The MAE loss used to $\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}), \bar{Y})$ is represented by Eq. (9):

$$\begin{aligned}\bar{\mathcal{L}}_{\mathrm{MAE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{Y}) = \frac{2^{K-|\bar{Y}|-1}}{2^K-2}\sum_{y=1, y\notin\bar{Y}}^{K}(1-f_y(\boldsymbol{x}))+ \\ \frac{2^{K-|\bar{Y}|-1}-1}{2^K-2}\sum_{y=1, y\notin\bar{Y}}^{K}f_y(\boldsymbol{x}) + \frac{2^{K-|\bar{Y}|}-1}{2^K-2}\sum_{y=1, y\in\bar{Y}}^{K}f_y(\boldsymbol{x}),\end{aligned} \tag{9}$$

where the gradient with respect to $\boldsymbol{\theta}$ is given by:

$$\frac{\partial\bar{\mathcal{L}}_{\mathrm{MAE}}}{\partial\boldsymbol{\theta}} = \begin{cases}-\frac{1}{2^K-2}\,\nabla_{\boldsymbol{\theta}}\, f_y(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } y\notin\bar{Y} \\ \frac{2^{K-|\bar{Y}|}-1}{2^K-2}\,\nabla_{\boldsymbol{\theta}}\, f_y(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } y\in\bar{Y}\end{cases}. \tag{10}$$

Differing from $\bar{\mathcal{L}}_{\mathrm{BCE}}$, $\bar{\mathcal{L}}_{\mathrm{MAE}}$ eliminates the impact of the prediction of $f_y(\boldsymbol{x};\boldsymbol{\theta})$ on its gradient and prevents an unbounded situation. Comparing the gradients of these two loss functions, $\bar{\mathcal{L}}_{\mathrm{MAE}}$ treats each example equally, while $\bar{\mathcal{L}}_{\mathrm{BCE}}$ implicitly assigns more weight to difficult examples to expedite convergence [19]. This indicates that $\bar{\mathcal{L}}_{\mathrm{MAE}}$ may facilitate stable optimization of our risk estimator but converge at a slower rate [39]. To balance convergence rate and stability, we improve $\bar{\mathcal{L}}_{\mathrm{MAE}}$ by proposing an upper-bound surrogate loss function $\bar{\mathcal{L}}'(\boldsymbol{x}, \bar{Y})$, which is defined as:

$$\begin{aligned}\bar{\mathcal{L}}'(\boldsymbol{x}, \bar{Y}) = &-\sum_{y=1, y\in\bar{Y}}^{K} e^{2^{1-|\bar{Y}|}}\log(1-f_y(\boldsymbol{x})) \\ &-\sum_{y=1, y\notin\bar{Y}}^{K} e^{-2^{-|\bar{Y}|}}\left\{\log(f_y(\boldsymbol{x}))+\log(1-f_y(\boldsymbol{x}))\right\}.\end{aligned} \tag{11}$$

The proof is presented in Appendix E. The gradient of $\bar{\mathcal{L}}'$ with respect to $\boldsymbol{\theta}$ is given by:

$$\frac{\partial\bar{\mathcal{L}}'}{\partial\boldsymbol{\theta}} = \begin{cases}\left(\frac{e^{-2^{-|\bar{Y}|}}}{1-f_y(\boldsymbol{x};\boldsymbol{\theta})} - \frac{e^{-2^{-|\bar{Y}|}}}{f_y(\boldsymbol{x};\boldsymbol{\theta})}\right)\nabla_{\boldsymbol{\theta}}\, f_y(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } y\notin\bar{Y} \\ e^{2^{1-|\bar{Y}|}}\frac{1}{1-f_y(\boldsymbol{x};\boldsymbol{\theta})}\,\nabla_{\boldsymbol{\theta}}\, f_y(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } y\in\bar{Y}\end{cases}. \tag{12}$$

Compared to $\bar{\mathcal{L}}_{\mathrm{MAE}}$, Eq. (12) clearly verifies that optimizing the upper-bound loss function $\bar{\mathcal{L}}'$ improves the convergence rate by implicitly assigning more weight to difficult examples. However, similar to $\bar{\mathcal{L}}_{\mathrm{BCE}}$, it is still sensitive to the prediction of $f_y(\boldsymbol{x};\boldsymbol{\theta})$ on non-CLs. Whether $f_y(\boldsymbol{x};\boldsymbol{\theta})$ is close to 0 or 1 for non-CLs, this can lead to either infinitely small or infinitely large gradients, resulting in an unstable learning process and hindering convergence. Hence, the model's predictions on non-CLs must maintain a lower confidence to prevent unbounded gradients when using $\bar{\mathcal{L}}'$. In fact, a model benefits from higher confidence in its predictions. For example, if a label $y\notin\bar{Y}$ is relevant to $\boldsymbol{x}$, the prediction of $f_y(\boldsymbol{x};\boldsymbol{\theta})$ should be close to 1; conversely, if it's irrelevant, $f_y(\boldsymbol{x};\boldsymbol{\theta})\rightarrow 0$. Unfortunately, the model trained with $\bar{\mathcal{L}}'$ in Eq. (11) cannot assign high-confidence predictions for non-CLs, as doing so would risk infinite gradients, thereby making training unstable.

Previous work addresses the issue by assigning higher predictions to any label belonging to non-CLs as relevant labels [16]. However, this approach ignores the fact that non-CLs consist of both relevant and irrelevant labels, which means it does not sufficiently handle the irrelevant labels. As observed in the part of Eq. (12) that calculates the gradients for non-CLs, $e^{-2^{-|\bar{Y}|}}/(1-f_y(\boldsymbol{x};\boldsymbol{\theta}))\rightarrow\infty$ when the predictions for non-CLs are close to 1. In fact, this factor assists gradient descent by encouraging the prediction of labels in non-CLs as irrelevant. Motivated by this observation, we propose a loss function called *confidence truncated loss* (CTL). By introducing a confidence truncation threshold, CTL prevents labels with high prediction confidence from computing $e^{-2^{-|\bar{Y}|}}/(1-f_y(\boldsymbol{x};\boldsymbol{\theta}))$, helping to alleviate the issue of unbounded gradients. The CTL is defined as:

$$\begin{aligned}\bar{\mathcal{L}}_{\mathrm{CTL}}(\boldsymbol{x}, \bar{Y}) = &-\sum_{y=1, y\in\bar{Y}}^{K} e^{2^{1-|\bar{Y}|}}\log(1-f_y(\boldsymbol{x})) \\ &-\sum_{y=1, y\notin\bar{Y}}^{K} e^{-2^{-|\bar{Y}|}}\left\{\log(f_y(\boldsymbol{x}))+\epsilon_{y,\lambda}\log(1-f_y(\boldsymbol{x}))\right\},\end{aligned} \tag{13}$$

where $\epsilon_{y,\lambda} = \mathbb{I}(f_y(\boldsymbol{x})\leq\lambda)$, with $\mathbb{I}(\cdot)$ denoting the indicator function. The threshold $\lambda\in[0,1]$ is used to truncate high-confidence predictions. This prevents Eq. (13) from comput-

Algorithm 1: Multiple CLs in MLCLL with CTL

```
Input:
$\bar{D}$: the multiple-complementary-label training set;
$E$: the number of epochs;
$\lambda$: the threshold;
$\mathcal{A}$: an external stochastic optimization algorithm;
Output:
$\boldsymbol{\theta}$: model parameter for $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$;
for epoch in $E$ do
    Let $\phi$ be the risk, $\phi = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathcal{L}}_{\mathrm{CTL}}(\boldsymbol{x}_i, \bar{Y}_i)$;
    Set gradient $- \bigtriangledown_{\boldsymbol{\theta}} \phi$;
    Update $\boldsymbol{\theta}$ by $\mathcal{A}$;
end
```

ing predictions close to 1 in $\log(1 - f_y(\boldsymbol{x}))$ for non-CLs, thereby avoiding infinite gradients and maintaining a stable learning process. The overall procedure of the proposed approach is shown in Algorithm 1.

5 Experiments

In this section, we evaluate the performance of CTL and the risk-consistent estimator with various loss functions. We adopt four common MLL criteria, including *one error*, *coverage*, *ranking loss*, and *average precision*, to evaluate the approaches' performance. A higher *average precision* value signifies superior performance, while smaller values for the other criteria indicate better performance. Our experiments are implemented using PyTorch [40] and NVIDIA RTX 3090 Ti. The code of this paper is available at https://github.com/gaoyi439/CTL.

5.1 Experimental Settings

Datasets & Pre-processing. We conduct experiments on seven MLL datasets[1]. Following prior work [4], [16], [41], we remove rare labels and their corresponding instances from datasets with more than 100 labels, and keep label spaces under 15. To generate multiple CLs, we first instantiate $p(s) = C_K^s/(2^K - 2)$, where $\forall s \in \{1, 2, \ldots, K-1\}$, $p(s)$ denotes the ratio of the number of label sets whose size is s to the number of all possible label sets. Next, for each instance $\boldsymbol{x}$, we randomly sample s from $p(s)$, and then uniformly sample a complementary label set $\bar{Y}$ with size s. Characteristics of each dataset are described through various statistics, including the number of features $dim(\mathcal{S})$, the number of instances $|\mathcal{S}|$, the number of possible labels $L(\mathcal{S})$, the average number of relevant labels per instance $LCard(\mathcal{S})$, and the number of CLs per instance avg.#CLs. The details are provided in Table 1.

Comparison Approaches. We adopt a MLL approach called CCMN [42] as a baseline, implemented by treating non-CLs ($\mathcal{Y} - \bar{Y}$) as relevant labels. As a comparison approach, we employ a PML approach, fpml [43], whose learning relies on treating non-CLs as candidate labels. L-UW [35] belongs to complementary label learning in a multi-class classification scenario, where we respectively use the Sigmoid layer and BCE loss to replace the Softmax layer and cross-entropy loss to make L-UW adapt to the problem setting of ML-MCL. Additionally, we compare with two MLCLL losses: GDF and MAE [16]. To verify the feasibility of our analysis in Section 4, we use BCE and MAE losses for our risk-consistent estimator, i.e. $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and $\bar{\mathcal{L}}_{\mathrm{MAE}}$, as baselines.

TABLE 1
Characteristics of datasets.

Datasets	$dim(\mathcal{S})$	$\|\mathcal{S}\|$	$L(\mathcal{S})$	$LCard(\mathcal{S})$	avg.#CLs
Corel16k	500	11153	153	1.77	7.48
tmc2007	981	28596	22	2.16	10.98
rcv1-s2	944	5252	101	1.67	7.44
rcv1-s3	944	5410	101	1.61	7.49
rcv1-s4	944	5761	101	1.48	7.55
scene	294	2407	6	1.07	2.96
VOC2007	3×448×448	9963	20	1.46	10.04

Setup. We employ SGD with a momentum of 0.9 for optimization, and $\lambda = 0.3$. The batch size and training epochs are set to 256 and 200, respectively. We set weight decay as 10^{-3}. The learning rate is chosen from $\{10^{-1}, 10^{-2}, 10^{-3}\}$, where the learning rate is reduced by a factor of 0.1 at 100-th and 150-th epochs [44]. Since the VOC2007 dataset comprises raw color images, we adopt an 18-layer ResNet as the predictive model, while the remaining datasets use a linear model for classification. We evaluate approaches over 5 trials for the VOC2007 dataset, while the other datasets undergo ten-fold cross-validation. Note that we apply the same model and hyper-parameters for all approaches except for fpml, since these approaches are implemented using neural networks. Here, the training data only involves CLs, while the test data are labeled with the sets of relevant labels to evaluate the performance of approaches. We report results as the mean and *standard deviation* (std) of four criteria, where $\downarrow$ / $\uparrow$ indicates that smaller/larger values of criteria are better performance.

5.2 Experimental Results

Results. Table 2 reports empirical results of four criteria across 7 datasets, where fpml is denoted as "-" in the VOC2007 dataset since fpml cannot handle raw images in the VOC2007 dataset. In Table 2, CTL achieves comparable performance against all baselines across most datasets, which demonstrates the effectiveness of the proposed CTL approach in the ML-MCL scenario. Notably, CTL outperforms CCMN and fpml across all datasets for four criteria, which proves its suitability for learning with multiple CLs over MLL or PML approaches. Compared to L-UW, the *average precision* of CTL on the scene dataset is 0.354 higher than that of L-UW. This performance gap is attributed to the reliance of complementary label learning approaches on the presence of one relevant label per instance in a multi-class scenario. Additionally, CTL is superior to MLCLL approaches on most datasets, which suggests that its design, considering the number of CLs per instance, is more effective than existing MLCLL approaches that focus on identifying labels belonging to non-CLs or CLs during the

[1] Publicly available at https://mulan.sourceforge.net/datasets-mlc.html

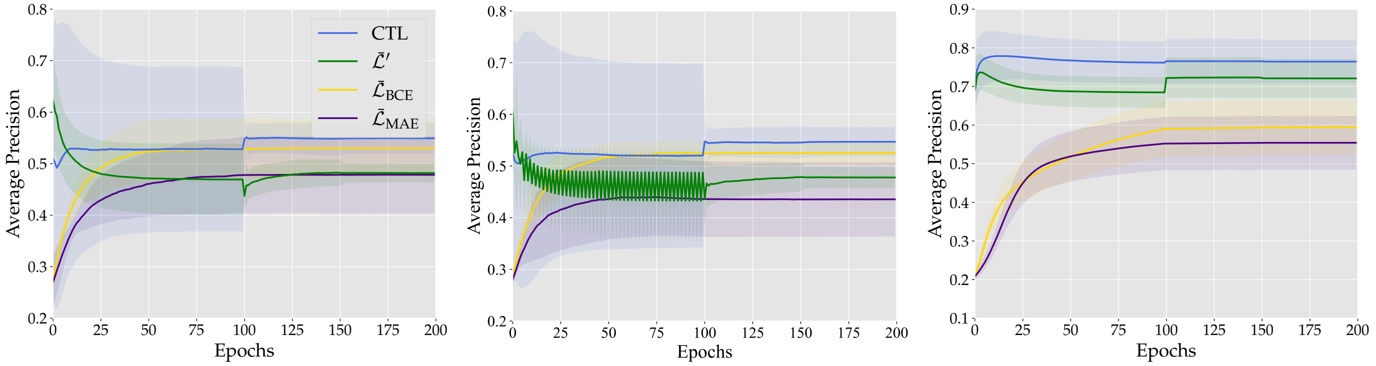
TABLE 2
Experimental results (mean±std) on 7 datasets. The best performance of each dataset is shown in **boldface**, where ●/○ indicates whether CTL is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

Approaches	CCMN	fpml	L-UW	GDF	MAE	$\bar{\mathcal{L}}_{\mathrm{BCE}}$	$\bar{\mathcal{L}}_{\mathrm{MAE}}$	CTL
One Error↓								
Corel16k	.753±.033●	.717±.057●	.708±.061●	.639±.048	.688±.064●	.688±.061●	.714±.059●	**.634±.049**
tmc2007	.404±.107●	.434±.098●	.433±.098●	.253±.089	.410±.101●	.420±.101●	.434±.098●	**.250±.088**
rcv1-s2	.476±.175●	.878±.039●	.800±.120●	.431±.135●	.517±.080●	.643±.086●	.774±.169●	**.422±.139**
rcv1-s3	.474±.150●	.861±.064●	.738±.230●	.440±.139●	.449±.094●	.640±.029●	.796±.160●	**.430±.141**
rcv1-s4	.533±.123●	.539±.137●	.742±.067●	.401±.131●	.507±.029●	.584±.040●	.768±.091●	**.388±.137**
scene	.691±.036●	.788±.019●	.779±.025●	.270±.026	.749±.031●	.349±.027●	.764±.019●	**.268±.029**
VOC2007	.595±.000●	-	.322±.035●	.523±.015●	.621±.025●	.201±.011●	.527±.009●	**.106±.007**
Coverage↓								
Corel16k	.480±.041●	.407±.062●	.390±.060●	.360±.036	.363±.042	.392±.059●	.396±.065●	**.358±.033**
tmc2007	.333±.018●	.324±.029●	.268±.028●	.170±.009●	.217±.026●	.307±.022●	.344±.023●	**.154±.008**
rcv1-s2	.295±.044●	.428±.057●	.271±.078●	.245±.095	**.204±.017**	.285±.047●	.297±.045●	.238±.095
rcv1-s3	.292±.054●	.430±.048●	.246±.036●	.234±.097	**.217±.064**	.282±.040●	.321±.049●	.232±.086
rcv1-s4	.322±.049●	.348±.084●	.232±.040●	.206±.090	**.170±.021**○	.256±.050●	.299±.047●	.203±.087
scene	.355±.024●	.401±.022●	.330±.024●	.093±.007●	.181±.011●	.117±.010●	.369±.017●	**.091±.008**
VOC2007	.427±.025●	-	.098±.010●	.208±.015●	.316±.029●	.094±.007●	.208±.011●	**.067±.009**
Ranking Loss↓								
Corel16k	.382±.036●	.312±.073●	.294±.067●	.265±.035	.267±.046	.298±.067●	.305±.073●	**.259±.035**
tmc2007	.194±.040●	.188±.054●	.152±.049●	.076±.021●	.119±.044●	.172±.047●	.197±.050●	**.067±.021**
rcv1-s2	.205±.064●	.344±.028●	.198±.051●	.165±.081●	**.132±.007**	.207±.031●	.214±.027●	.158±.082
rcv1-s3	.209±.070●	.351±.027●	.181±.024	.166±.083	.165±.031	.209±.021●	.243±.040●	**.164±.074**
rcv1-s4	.252±.071●	.240±.115●	.181±.019●	.151±.073	**.117±.009**○	.197±.028●	.241±.026●	.149±.070
scene	.407±.031●	.463±.025●	.377±.028●	.094±.009	.202±.012●	.123±.012●	.420±.019●	**.092±.010**
VOC2007	.361±.022●	-	.064±.008●	.166±.013●	.252±.026●	.058±.005●	.162±.010●	**.032±.006**
Average Precision↑								
Corel16k	.376±.028●	.422±.055●	.436±.057●	.484±.037	.457±.051●	.444±.054●	.424±.055●	**.491±.038**
tmc2007	.581±.068●	.553±.070●	.587±.073●	.771±.059	.626±.074●	.595±.071●	.554±.069●	**.781±.059**
rcv1-s2	.578±.099	.340±.019●	.464±.030●	.646±.111●	.624±.024●	.533±.057●	.479±.074●	**.656±.115**
rcv1-s3	.582±.095●	.345±.033●	.497±.099●	.637±.110	.631±.022	.526±.023●	.444±.076●	**.645±.108**
rcv1-s4	.538±.071●	.532±.110●	.503±.032●	.677±.116●	.656±.018●	.561±.037●	.461±.051●	**.686±.118**
scene	.519±.024●	.443±.018●	.486±.024●	.838±.015	.590±.018●	.791±.017●	.473±.016●	**.840±.016**
VOC2007	.399±.012●	-	.756±.023●	.557±.018●	.458±.019●	.822±.009●	.566±.013●	**.902±.006**

learning process. Furthermore, CTL demonstrates significant improvement compared to $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and $\bar{\mathcal{L}}_{\mathrm{MAE}}$, which validates the feasibility of our analysis to improve the risk-consistent estimator.

Effect of CTL. Fig. 1 describes the curves of *average precision* for CTL, $\bar{\mathcal{L}}'$ (i.e., CTL without the confidence truncated threshold λ), $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and $\bar{\mathcal{L}}_{\mathrm{MAE}}$ across 200 epochs. As observed in Fig. 1, the curve of $\bar{\mathcal{L}}_{\mathrm{MAE}}$ is inferior than that of $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and CTL, which supports the analysis in Section 4 that $\bar{\mathcal{L}}_{\mathrm{MAE}}$ remains stable during the learning process but exhibits a lower convergence rate [16], [45]. This finding confirms that choosing an upper bounded of $\bar{\mathcal{L}}_{\mathrm{MAE}}$ as the optimization objective can enhance convergence rate. Furthermore, we notice fluctuations in the curve of $\bar{\mathcal{L}}'$ during the learning process, especially in the rcv1-s3 dataset. This fluctuation stems from the inconsistent predictions for non-CLs, where $\bar{\mathcal{L}}'$ encourages assigning high-confidence predictions of the same label as an irrelevant label and a relevant label for an instance, which leads to an unstable learning process and challenging convergence. In contrast, the curve of CTL exhibits a stable learning process, which indicates that a confidence truncated threshold indeed prevents fluctuations, thus improving gradient updates to converge stably during the learning process.

Ablation Studies. To validate the contributions of two main strategies for CTL, we compare CTL with three variants: (1) *CTL w/o upper bound*: This variant employs $\bar{\mathcal{L}}_{\mathrm{MAE}}$ with the confidence truncated threshold λ to learn a classifier, i.e., using the equation $\frac{2^{K-|\bar{Y}|-1}}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}(1-f_y(\boldsymbol{x}))+\frac{2^{K-|\bar{Y}|-1}-1}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}\epsilon_{y,\lambda}f_y(\boldsymbol{x})+\frac{2^{K-|\bar{Y}|}-1}{2^K-2}\sum_{y=1,y\in\bar{Y}}^{K}f_y(\boldsymbol{x})$ to learn; (2) *CTL w/o* λ: This variant removes the confidence truncated threshold λ from CTL, i.e., using only $\bar{\mathcal{L}}'$ for learning; (3) *CTL w/o upper bound* & λ: This variant indicates CTL without the strategies of upper bound and the confidence truncated threshold λ, which employs $\bar{\mathcal{L}}_{\mathrm{MAE}}$ for training. Table 3 displays four criteria for these three variants and CTL across 7 datasets. As can be seen from Table 3, variant



(a) Rcv1-s2

(b) Rcv1-s3

(c) Tmc2007

Fig. 1. *Average precision* on various datasets. Dark colors represent the mean testing results, while light colors correspond to the std.

TABLE 3
Ablation studies. The best results (mean±std) for each dataset are shown in **boldface**.

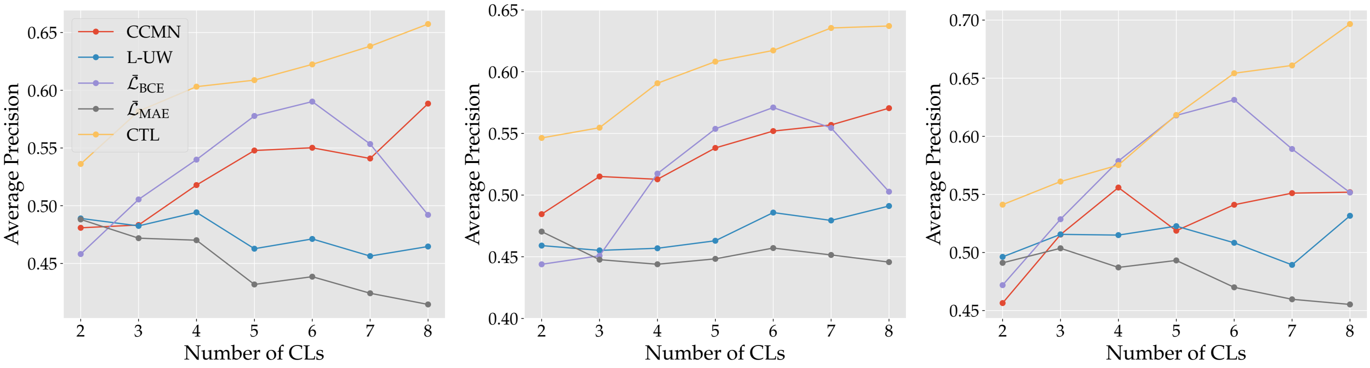
Datasets	Corel16k	tmc2007	rcv1-s2	rcv1-s3	rcv1-s4	scene	VOC2007
			One Error↓				
CTL	**.634±.049**	**.250±.088**	**.422±.139**	**.430±.141**	**.388±.137**	**.268±.029**	**.106±.007**
CTL w/o upper bound	.769±.066	.702±.061	.834±.173	.843±.161	.827±.142	.835±.021	.595±.000
CTL w/o λ	.652±.039	.281±.076	.490±.170	.472±.130	.469±.074	.282±.024	.254±.009
CTL w/o upper bound & λ	.714±.059	.434±.098	.774±.169	.796±.160	.768±.091	.764±.019	.527±.009
			Coverage↓				
CTL	**.358±.033**	**.154±.008**	.238±.095	.232±.086	**.203±.087**	**.091±.008**	**.067±.009**
CTL w/o upper bound	.399±.052	.358±.019	.312±.028	.326±.029	.301±.043	.374±.017	.369±.094
CTL w/o λ	.370±.030	.190±.009	**.233±.062**	**.230±.061**	.213±.054	.095±.007	.133±.015
CTL w/o upper bound & λ	.396±.065	.344±.023	.297±.045	.321±.049	.299±.047	.369±.017	.208±.011
			Ranking Loss↓				
CTL	**.259±.035**	**.067±.021**	.158±.082	.164±.074	**.149±.070**	**.092±.010**	**.032±.006**
CTL w/o upper bound	.302±.056	.211±.046	.234±.009	.253±.020	.245±.022	.431±.017	.306±.085
CTL w/o λ	.274±.028	.089±.017	**.157±.050**	**.162±.048**	.158±.038	.098±.008	.087±.011
CTL w/o upper bound & λ	.305±.073	.197±.050	.214±.027	.243±.040	.241±.026	.420±.019	.162±.010
			Average Precision↑				
CTL	**.491±.038**	**.781±.059**	**.656±.115**	**.645±.108**	**.686±.118**	**.840±.016**	**.902±.006**
CTL w/o upper bound	.404±.051	.497±.063	.414±.055	.394±.060	.411±.057	.441±.013	.440±.050
CTL w/o λ	.471±.031	.739±.047	.627±.099	.630±.083	.638±.063	.832±.013	.821±.015
CTL w/o upper bound & λ	.424±.055	.554±.069	.479±.074	.444±.076	.461±.051	.473±.016	.566±.013

(1) is inferior to variant (3) on almost all cases, which presents that the confidence truncated threshold strategy is not suitable for $\bar{\mathcal{L}}_{\text{MAE}}$ and even leads to negative effects. This is because $\bar{\mathcal{L}}_{\text{MAE}}$ does not struggle in the trouble of infinite gradients, and applying a confidence truncated threshold strategy will affect its performance. On the other hand, variant (2) surpasses variant (3) on all datasets, which proves that optimizing an upper bound of $\bar{\mathcal{L}}_{\text{MAE}}$ benefits gradient updates. Moreover, CTL outperforms variant (2) on almost all datasets, which demonstrates that the confidence truncated threshold strategy can prevent an unstable convergence process in variant (2) ($\bar{\mathcal{L}}'$). In summary, these two strategies enable CTL to benefit from gradient updates and improve risk-consistent estimator.

Performance of CTL with Different Number of CLs. Additionally, we explore the performance impact when the size of the complementary label set s is fixed for each instance, with CLs randomly sampled from $\mathcal{Y} - Y$. We conduct experiments with varying values of s for each instance to investigate how the number of CLs influences the performance of approaches. Table 4 presents the results of four criteria for various approaches on the rcv1-s2 and VOC2007 datasets when $s \in \{2, 3\}$. Meanwhile, Fig. 2 illustrates the *average precision* of different approaches across three datasets — rcv1-s2, rcv1-s3, and rcv1-s4 — ranging from $s = 2$ to $s = 8$. The experimental results in Table 4 and Fig. 2 reveal that the performance of the CTL approach consistently improves as the number of CLs per instance increases. Specifically, when comparing the results for $s = 2$ and $s = 3$ on the rcv1-s2 and VOC2007 datasets, the

TABLE 4
Experimental results (mean±std) on the training data with a fixed number of CLs. The best performance of each dataset is shown in **boldface**, where ●/○ denotes whether CTL is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

CLs	Datasets	CCMN	L-UW	$\bar{\mathcal{L}}_{\mathrm{BCE}}$	$\bar{\mathcal{L}}_{\mathrm{MAE}}$	CTL
			One Error↓			
s=2	rcv1-s2	.611±.126●	.690±.050●	.676±.049●	.681±.069●	**.579±.164**
	VOC2007	.718±.172●	.924±.010●	.595±.000●	.573±.036●	**.449±.111**
s=3	rcv1-s2	.597±.167●	.713±.059●	.624±.035●	.743±.073●	**.500±.187**
	VOC2007	.596±.002●	.916±.010●	.397±.077●	.518±.018●	**.252±.081**
			Coverage↓			
s=2	rcv1-s2	.390±.030●	.325±.020	.363±.054●	.338±.044●	**.316±.088**
	VOC2007	.428±.025●	.519±.035●	.365±.003●	**.268±.011**	.272±.086
s=3	rcv1-s2	.375±.037●	.301±.033	.324±.046●	.302±.046	**.298±.124**
	VOC2007	.453±.032●	.509±.034●	.205±.057	.227±.019●	**.202±.027**
			Ranking Loss↓			
s=2	rcv1-s2	.304±.050●	.235±.015	.285±.039●	.243±.026●	**.232±.079**
	VOC2007	.359±.027●	.475±.036●	.296±.002●	**.214±.011**	.216±.081
s=3	rcv1-s2	.290±.052●	.222±.022	.246±.031●	.222±.031	**.218±.111**
	VOC2007	.381±.029●	.464±.035●	.154±.049	.177±.017●	**.148±.026**
			Average Precision↑			
s=2	rcv1-s2	.481±.081●	.489±.029●	.458±.049●	.488±.028●	**.536±.118**
	VOC2007	.355±.058●	.208±.017●	.433±.000●	.510±.027●	**.587±.116**
s=3	rcv1-s2	.483±.090●	.483±.020●	.505±.034●	.472±.022●	**.582±.154**
	VOC2007	.390±.018●	.216±.017●	.641±.078●	.561±.020●	**.755±.061**



(a) Rcv1-s2 (b) Rcv1-s3 (c) Rcv1-s4

Fig. 2. *Average precision* on various datasets with different number of CLs. Dark colors represent the mean testing results, while light colors correspond to the std.

outcomes for $s = 3$ outperform those for $s = 2$ across all criteria, with notable improvements in metrics such as *coverage*, *ranking loss*, and *average precision*. This indicates that a greater number of CLs per instance can provide more supervision information for learning, thus boosting the model's performance. The representation in Fig. 2 further corroborates these findings, showing a clear upward trend in CTL's performance as s increases. The analysis underscores the importance of considering the number of CLs in MLCLL algorithms and highlights CTL as a robust approach that effectively utilizes more label information to achieve improved learning outcomes. Furthermore, CTL achieves promising performance against other baselines, illustrating the effectiveness and flexibility of CTL as it can adapt to varying definitions of s.

Effect of Various λ on CTL. Here, we investigate the effect of various confidence truncated thresholds λ on CTL performance. Based on the analysis in Section 4, where a large threshold is deemed meaningless, we select λ from the set $\{0.1, 0.3, 0.5, 0.6\}$. The results presented in Table 5.2 indicate that CTL achieves optimal performance when $\lambda = 0.3$, outperforming other threshold values in most cases. This optimal threshold enables *average precision* to peak at $\lambda = 0.3$, indicating that this threshold enhances the model's

TABLE 5
Results with various λ. The best performance for each dataset is shown in **boldface**.

λ	0.1	0.3	0.5	0.6
		One Error↓		
Corel16k	.638±.049	.634±.049	**.627±.054**	.629±.057
tmc2007	.252±.086	**.250±.088**	.258±.096	.268±.098
rcv1-s4	.397±.132	**.388±.137**	.396±.103	.402±.079
scene	.268±.028	.268±.029	**.267±.029**	.270±.026
VOC2007	.118±.006	.106±.007	**.104±.004**	.110±.006
		Coverage↓		
Corel16k	.356±.033	**.358±.033**	.366±.034	.377±.033
tmc2007	.168±.009	**.154±.008**	.183±.010	.194±.005
rcv1-s4	.209±.088	.203±.087	**.201±.072**	.209±.062
scene	.092±.007	**.091±.008**	.093±.008	.095±.007
VOC2007	.077±.006	**.067±.009**	.072±.000	.074±.001
		Ranking Loss↓		
Corel16k	.259±.034	**.259±.035**	.263±.034	.274±.033
tmc2007	.075±.022	**.067±.021**	.083±.023	.089±.021
rcv1-s4	.154±.070	.149±.070	**.146±.054**	.152±.045
scene	.094±.009	**.092±.010**	.094±.010	.096±.009
VOC2007	.040±.004	**.032±.006**	.036±.000	.038±.002
		Average Precision↑		
Corel16k	.486±.037	.491±.038	**.493±.041**	.486±.041
tmc2007	.771±.059	**.781±.059**	.761±.064	.746±.064
rcv1-s4	.678±.115	**.686±.118**	.684±.090	.676±.075
scene	.839±.015	**.840±.016**	.839±.016	.836±.015
VOC2007	.888±.003	**.902±.006**	.898±.003	.894±.003

precision in identifying instances with high confidence. The consistency of this finding across diverse datasets, such as Corel16k, tmc2007, rcv1-s4, scene, and VOC2007, suggests that the choice of $\lambda = 0.3$ generalizes well to diverse data distributions. This robustness emphasizes the importance of selecting an appropriate λ value that balances the trade-off between discarding overly confident predictions and retaining valuable information for learning. The consistently superior performance at $\lambda = 0.3$ in most cases solidifies our decision to set λ to 0.3 for all experiments. This choice ensures that CTL operates at its peak, providing the best predictions possible.

Execution Time. In Table 6, we present the running time of each approach on the six datasets. Shorter execution times generally indicate a lower computational complexity of the approach. It can be observed from Table 6 that the execution times of CTL exhibit slightly higher running times than some baselines, such as L-UW and GDF, but the differences remain small. Specifically, CTL shows marginally longer times in the tmc2007 dataset (7.73×10^2 seconds versus L-UW's 7.16×10^2 seconds), the relative difference remains within 8% — a reasonable trade-off considering the significant performance improvements shown in Table 2. In particular, the absolute time difference (57 seconds) is negligible for real-world applications that prioritize model performance. This demonstrates that CTL achieves a favorable balance between computational efficiency and effectiveness.

TABLE 6
The running time (in 10^2 seconds) of each approach.

Datasets	CCMN	L-UW	MAE	GDF	$\bar{\mathcal{L}}_{\mathrm{BCE}}$	$\bar{\mathcal{L}}_{\mathrm{MAE}}$	CTL
Corel16k	4.70	4.05	4.22	4.13	4.66	4.35	4.31
tmc2007	8.63	7.16	7.45	7.34	7.55	7.43	7.73
rcv1-s2	3.47	3.03	3.13	3.01	2.96	2.95	3.12
rcv1-s3	4.36	3.10	3.24	3.05	3.00	2.96	3.15
rcv1-s4	3.63	3.13	3.34	3.13	3.04	3.04	3.21
scene	2.82	2.49	2.56	2.46	2.36	2.39	2.45

6 Conclusion

In this paper, we propose a novel problem setting, called ML-MCL, which expands the MLCLL problem to learn with multiple CLs, while facing greater challenges due to the uncertain number of CLs. To solve this problem, we theoretically derive a risk-consistent estimator with an estimation error bound at $\mathcal{O}(1/\sqrt{n})$ convergence rate by analyzing the process of generating multiple CLs. Although our risk estimator does not depend on specific models or loss functions, the risk estimator may produce unbounded gradients when using certain loss functions, which can lead to an unstable learning process and challenging convergence. Therefore, we design CTL to improve the risk-consistent estimator to prevent the above issues. Extensive experiments validate the effectiveness of the proposed approaches. It is noteworthy that the effectiveness of ML-MCL heavily depends on the quality of CLs. Errors in CLs could distort the estimated risk, resulting in a less accurate classifier compared to a scenario without such errors. Addressing this issue effectively would require a comprehensive solution, encompassing an alternative data generation process and a tailored loss function, which is beyond the scope of our current study. Therefore, we intend to tackle these error-related challenges in our future work.

References

[1] M.-L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, 2015.

[2] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014.

[3] Y. Zhang and M. Zhang, "Generalization analysis for multi-label learning," in *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, 2024.

[4] Y. Gao, M. Xu, and M.-L. Zhang, "Complementary to multiple labels: A correlation-aware correction approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] X. Zhang and T. Luo, "Imbalanced multi-instance multi-label learning via tensor product-based semantic fusion," *Frontiers of Computer Science*, vol. 19, no. 8, p. 198346, 2025.

[6] F. Sun, M. Xie, and S. Huang, "A deep model for partial multi-label image classification with curriculum-based disambiguation," *Machine Intelligence Research*, vol. 21, no. 4, pp. 801–814, 2024.

[7] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[8] N. Xu, C. Qiao, J. Lv, X. Geng, and M. Zhang, "One positive label is sufficient: Single-positive multi-label learning with label enhancement," in *Advances in Neural Information Processing Systems 35*, New Orleans, LA, 2022.

[9] E. Cole, O. M. Aodha, T. Lorieul, P. Perona, D. Morris, and N. Jojic, "Multi-label learning from single positive labels," in *Proceedings of 2021 IEEE Conference on Computer Vision and Pattern Recognition*, Virtual Event, 2021, pp. 933–942.

[10] X. Niu, H. Han, S. Shan, and X. Chen, "Multi-label co-regularization for semi-supervised facial action unit recognition," in *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019, pp. 907–917.

[11] G. Lin, K. Liao, B. Sun, Y. Chen, and F. Zhao, "Dynamic graph fusion label propagation for semi-supervised multi-modality classification," *Pattern Recognit.*, vol. 68, pp. 14–23, 2017.

[12] A. Kanehira and T. Harada, "Multi-label ranking from positive and unlabeled data," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 5138–5146.

[13] C. Hsieh, N. Natarajan, and I. S. Dhillon, "PU learning for matrix completion," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Lille, France, 2015, pp. 2445–2453.

[14] J.-Y. Hang and M.-L. Zhang, "Partial multi-label learning via label-specific feature corrections," *Science China Information Sciences*, vol. 68, no. 3, p. 132104, 2025.

[15] B. Li, Y. Zheng, B. Jin, T. Xiang, H. Wang, and L. Feng, "Asyco: an asymmetric dual-task co-training model for partial-label learning," *Science China Information Sciences*, vol. 68, no. 5, p. 152101, 2025.

[16] Y. Gao, M. Xu, and M. Zhang, "Unbiased risk estimator to multi-labeled complementary label learning," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, Macao, China, 2023, pp. 3732–3740.

[17] M. Rezaei, H. Yang, and C. Meinel, "Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15 329–15 348, 2020.

[18] Y. Tian and H. Jiang, "Recent advances in complementary label learning," *Information Fusion*, p. 102702, 2024.

[19] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual Event, 2020, pp. 3072–3081.

[20] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 191–202, 2018.

[21] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine learning*, vol. 73, no. 2, pp. 133–153, 2008.

[22] Y.-C. Li, Y. Song, and J.-B. Luo, "Improving pairwise ranking for multi-label image classification," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, Honolulu, HI, 2017, pp. 1837–1845.

[23] S. Burkhardt and S. Kramer, "Online multi-label dependency topic models for text classification," *Mach. Learn.*, vol. 107, no. 5, pp. 859–886, 2018.

[24] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[25] Y.-C. Li, Y. Song, and J.-B. Luo, "Improving pairwise ranking for multi-label image classification," in *Proceedings of 2017 IEEE conference on computer vision and pattern recognition*, Honolulu, HI, 2017, pp. 3617–3625.

[26] W. Gerych, T. Hartvigsen, L. Buquicchio, E. Agu, and E. A. Rundensteiner, "Recurrent bayesian classifier chains for exact multi-label classification," in *Advances in Neural Information Processing Systems 34*, Virtual Event, 2021, pp. 15 981–15 992.

[27] W.-T. Zhao, S.-F. Kong, J.-W. Bai, D. Fink, and C. P. Gomes, "HOT-VAE: learning high-order label correlation for multi-label classification via attention-based variational autoencoders," in *Proceedings of 35th AAAI Conference on Artificial Intelligence*, Virtual Event, 2021, pp. 15 016–15 024.

[28] N. Xu, Y. Liu, and X. Geng, "Partial multi-label learning with label distribution," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020, pp. 6510–6517.

[29] L. Sun, S. Feng, J. Liu, G. Lyu, and C. Lang, "Global-local label correlation for partial multi-label learning," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2021.

[30] Y. Tang, Y. Gao, Y. Luo, J. Yang, M. Xu, and M. Zhang, "Unlearning from weakly supervised learning," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, Jeju, South Korea, 2024, pp. 5000–5008.

[31] T. Ishida, G. Niu, W.-H. Hu, and M. Sugiyama, "Learning from complementary labels," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, 2017, pp. 5639–5649.

[32] T. Ishida, G. Niu, A. K. Menon, and M. Sugiyama, "Complementary-label learning for arbitrary losses and models," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, Long Beach, CA, 2019, pp. 2971–2980.

[33] X.-Y. Yu, T.-L. Liu, M.-M. Gong, and D.-C. Tao, "Learning with biased complementary labels," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, 2018, pp. 69–85.

[34] Y. Xu, M. Gong, J. Chen, T. Liu, K. Zhang, and K. Batmanghelich, "Generative-discriminative complementary learning," in *The 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020, pp. 6526–6533.

[35] Y. Gao and M.-L. Zhang, "Discriminative complementary-label learning with weighted loss," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, 2021, pp. 3587–3597.

[36] L. Feng, J. Lv, B. Han, M. Xu, G. Niu, X. Geng, B. An, and M. Sugiyama, "Provably consistent partial-label learning," in *Advances in Neural Information Processing Systems 33*, Virtual Event, 2020.

[37] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 2233–2241.

[38] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 2111, pp. 224–240, 2002.

[39] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems 31*, Montréal, Canada, 2018, pp. 8792–8802.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z.-M. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J.-J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019, pp. 8024–8035.

[41] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, 2018, pp. 4302–4309.

[42] M. Xie and S. Huang, "CCMN: A general framework for learning with class-conditional multi-label noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 154–166, 2023.

[43] G.-X. Yu, X. Chen, C. Domeniconi, J. Wang, Z. Li, Z.-L. Zhang, and X.-D. Wu, "Feature-induced partial multi-label learning," in *Proceedings of 2018 IEEE International Conference on Data Mining*, Singapore, 2018, pp. 1398–1403.

[44] D. Wang, L. Feng, and M. Zhang, "Learning from complementary labels via partial-output consistency regularization," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. Virtual Event: ijcai.org, 2021, pp. 3075–3081.

[45] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 1919–1925.

[46] C. McDiarmid *et al.*, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[47] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[48] A. Maurer, "A vector-contraction inequality for rademacher complexities," in *Algorithmic Learning Theory*, 2016.

Yi Gao received the PhD degree from Southeast University, China, in 2025. She obtained the BSc and MSc degrees in computer science from Northwest University, China, in 2017 and 2020 respectively. Currently, she is an assistant researcher at the School of Computer Science and Engineering, Southeast University, China. Her main research interests include machine learning and data mining, with a focus on learning from complementary labels.

Jing-Yi Zhu received the BSc degree in information management and information system from Central South University in 2023. She is currently pursuing the MSc degree in computer technology in Southeast University. Her main research interests include machine learning and data mining, with a focus on learning from complementary labels.

Miao Xu is a senior lecturer in the School of Electrical Engineering and Computer Science at the University of Queensland, Australia. She was awarded the Australian Research Council Discovery Early Career Researcher Award (DECRA) in 2023. Dr Xu specializes in machine learning and data mining, particularly focusing on the challenges of learning from imperfect information. Dr Xu earned a PhD from Nanjing University, where research efforts led to notable recognitions including the CAAI Outstanding Doctoral Dissertation Award.

Min-Ling Zhang received the BSc, MSc, and PhD degrees in computer science from Nanjing University, China, in 2001, 2004 and 2007, respectively. Currently, he is a Professor at the School of Computer Science and Engineering, Southeast University, China. His main research interests include machine learning and data mining. In recent years, Dr. Zhang has served as the General Co-Chairs of ACML'18, Program Co-Chairs of CCML'25, PAKDD'19, CCF-ICAI'19, ACML'17, CCFAI'17, PRICAI'16, Senior PC member or Area Chair of KDD 2021-2024, AAAI 2022-2025, IJCAI 2017-2024, ICML 2024, ICLR 2024, etc. He is also on the editorial board of IEEE Transactions on Pattern Analysis and Machine Intelligence, Science China Information Sciences, ACM Transactions on Intelligent Systems and Technology, Frontiers of Computer Science, Machine Intelligence Research, etc. Dr. Zhang is the Steering Committee Member of ACML and PAKDD, Vice-Chair of the CAAI (Chinese Association of Artificial Intelligence) Machine Learning Society. He is a Distinguished Member of CCF, CAAI, and Senior Member of AAAI, ACM, IEEE.

Appendix A
The Proof of Theorem 1

Theorem 1. *$\bar{p}(\boldsymbol{x}, \bar{Y})$ is a valid probability distribution, which satisfies non-negativity and $\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{Y})}[1] = 1$.*

Proof. Let $\bar{\mathcal{Y}}_j := \{\bar{Y} | \bar{Y} \in \bar{\mathcal{Y}}, |\bar{Y}| = j\}$ be all possible sets with a size of j. To maintain the validity of the problem setting, Y cannot be $\emptyset$ or $\mathcal{Y}$. So, let $\mathcal{Y}' = \{2^{\mathcal{Y}} - \emptyset - \mathcal{Y}\}$, and we have

$$\begin{aligned}
\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{Y})}[1] &= \int_{\mathcal{X}} \int_{\bar{\mathcal{Y}}} \bar{p}(\boldsymbol{x}, \bar{Y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\bar{Y} \\
&= \int_{\mathcal{X}} \sum_{\bar{Y} \in \bar{\mathcal{Y}}} \bar{p}(\boldsymbol{x}, \bar{Y}) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{\bar{Y} \in \bar{\mathcal{Y}}} \sum_{j=1}^{K-1} p(s=j) \bar{p}(\boldsymbol{x}, \bar{Y} | s=j) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{\bar{Y} \in \bar{\mathcal{Y}}} \sum_{j=1}^{K-|Y|} \sum_{Y \in \mathcal{Y}', Y \cap \bar{Y} = \emptyset} \frac{p(\boldsymbol{x}, Y)}{C_{K-|Y|}^{j}} p(s=j) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{Y \in \mathcal{Y}'} \sum_{j=1}^{K-|Y|} \sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \frac{p(\boldsymbol{x}, Y)}{C_{K-|Y|}^{j}} p(s=j) \mathrm{d}\boldsymbol{x} \quad (\because \bar{\mathcal{Y}}_j := \{\bar{Y} | \bar{Y} \in \bar{\mathcal{Y}}, |\bar{Y}| = j\}) \\
&= \int_{\mathcal{X}} \sum_{Y \in \mathcal{Y}'} \sum_{j=1}^{K-|Y|} p(\boldsymbol{x}, Y) p(s=j) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}'} p(\boldsymbol{x}, Y) \mathrm{d}\boldsymbol{x} \mathrm{d}Y \\
&= 1,
\end{aligned}$$

which concludes the proof of Theorem 1. □

Appendix B
The Proof of Lemma 2

Lemma 2. *Let $\bar{\mathcal{Y}}_j = \{\bar{Y} | \bar{Y} \in \bar{\mathcal{Y}}, |\bar{Y}| = j\}$. With Eq.* (2),

$$p(\boldsymbol{x}, Y) = \frac{1}{2^K - 2} \sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \bar{p}(\boldsymbol{x}, \bar{Y} | s=j).$$

Proof. With Eq. (2), we have

$$\begin{aligned}
\sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \bar{p}\left(\boldsymbol{x}, \bar{Y} | s=j\right) &= \sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \sum_{Y \in \mathcal{Y}', Y \cap \bar{Y} = \emptyset} \frac{1}{C_{K-|Y|}^{j}} p\left(\boldsymbol{x}, Y\right) \\
&= \sum_{Y \in \mathcal{Y}'} \sum_{\bar{Y} \in \bar{Y}_j, Y \cap \bar{Y} = \emptyset} \frac{1}{C_{K-|Y|}^{j}} p\left(\boldsymbol{x}, Y\right) \\
&= \sum_{Y \in \mathcal{Y}'} p\left(\boldsymbol{x}, Y\right) \qquad \left(\because |\mathcal{Y}'| = 2^K - 2\right) \\
&= \left(2^K - 2\right) p\left(\boldsymbol{x}, Y\right).
\end{aligned}$$

Hence, we can get $p(\boldsymbol{x}, Y) = \frac{1}{2^K-2} \sum_{\bar{Y} \in \bar{\mathcal{Y}}_j, Y \cap \bar{Y} = \emptyset} \bar{p}(\boldsymbol{x}, \bar{Y} | s=j)$. □

Appendix C
The Proof of Theorem 3

Theorem 3. *Under Lemma 2, $R(\boldsymbol{f}) = \bar{R}(\boldsymbol{f})$ based on the definitions of $\bar{p}(\boldsymbol{x}, \bar{Y})$ and $R(\boldsymbol{f})$. $\bar{R}(\boldsymbol{f})$ is expressed as:*

$$\bar{R}(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s=j) \bar{R}_j(\boldsymbol{f}),$$

where $\bar{R}_j(\boldsymbol{f}) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{Y}|s=j)} \left[\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}), \bar{Y})\right]$ and $\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}), \bar{Y}) = \frac{2^{K-j-1}}{2^K-2} \sum_{y=1, y \notin \bar{Y}}^{K} \ell_y(\boldsymbol{x}) + \frac{2^{K-j-1}-1}{2^K-2} \sum_{y=1, y \notin \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x}) + \frac{2^{K-j}-1}{2^K-2} \sum_{y=1, y \in \bar{Y}}^{K} \bar{\ell}_y(\boldsymbol{x})$.

Proof. According to the definition of $R(\boldsymbol{f})$, we have

$$R\left(\boldsymbol{f}\right) = \mathbb{E}_{p(\boldsymbol{x},Y)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] = \sum_{j=1}^{K-|\bar{Y}|} p\left(s=j\right)\mathbb{E}_{p(\boldsymbol{x},Y|s=j)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right].$$

For $\mathbb{E}_{p(\boldsymbol{x},Y|s=j)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right]$, we can obtain

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x},Y|s=j)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] &= \mathbb{E}_{p(\boldsymbol{x}|s=j)}\mathbb{E}_{p(Y|\boldsymbol{x},s=j)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}|s=j)}\left[\sum_{Y\in\mathcal{Y}'} p\left(Y|\boldsymbol{x},s=j\right)\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}|s=j)}\left[\sum_{Y\in\mathcal{Y}'}\sum_{j=1}^{K-|Y|} p\left(Y|\boldsymbol{x},s=j\right)p\left(s=j\right)\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}|s=j)}\left[\sum_{Y\in\mathcal{Y}'}\frac{1}{2^K-2}\sum_{\bar{Y}\in\bar{\mathcal{Y}}_j,Y\cap\bar{Y}=\emptyset} p\left(\bar{Y}|\boldsymbol{x},s=j\right)\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}|s=j)}\mathbb{E}_{p\left(\bar{Y}|\boldsymbol{x},s=j\right)}\left[\sum_{Y\in\mathcal{Y}',Y\cap\bar{Y}=\emptyset}\frac{1}{2^K-2}\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{\bar{p}\left(\boldsymbol{x},\bar{Y}|s=j\right)}\left[\sum_{Y\in\mathcal{Y}',Y\cap\bar{Y}=\emptyset}\frac{1}{2^K-2}\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \mathbb{E}_{\bar{p}\left(\boldsymbol{x},\bar{Y}|s=j\right)}\left[\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y})\right] \\
&= \bar{R}_j(\boldsymbol{f}).
\end{aligned}$$

The fourth equality above is derived from Lemma 2. The specific reasoning is as follows:

$$\begin{aligned}
p(\boldsymbol{x},Y) = \sum_{j=1}^{K-|Y|} p\left(\boldsymbol{x},Y|s=j\right)p\left(s=j\right) &= \frac{1}{2^K-2}\sum_{\bar{Y}\in\bar{\mathcal{Y}}_j,Y\cap\bar{Y}=\emptyset}\bar{p}\left(\boldsymbol{x},\bar{Y}|s=j\right) \\
\Rightarrow \sum_{j=1}^{K-|Y|} p(s=j)p(\boldsymbol{x}|s=j)p(Y|\boldsymbol{x},s=j) &= \frac{1}{2^K-2}\sum_{\bar{Y}\in\bar{\mathcal{Y}}_j,Y\cap\bar{Y}=\emptyset} p(\bar{Y}|\boldsymbol{x},s=j)p(\boldsymbol{x}|s=j).
\end{aligned}$$

Since the generation of s is independent of $\boldsymbol{x}$, so $p\left(\boldsymbol{x}|s=j\right) = p\left(\boldsymbol{x}\right)$. Then, we have

$$\sum_{j=1}^{K-|Y|} p\left(Y|\boldsymbol{x},s=j\right)p\left(s=j\right) = \frac{1}{2^K-2}\sum_{\bar{Y}\in\bar{\mathcal{Y}}_j,Y\cap\bar{Y}=\emptyset} p\left(\bar{Y}|\boldsymbol{x},s=j\right).$$

According to the definition of Eq. (2), $\bar{p}\left(\boldsymbol{x},\bar{Y}|s=j\right) = 0$ for each $j > K-|Y|$, we have

$$\bar{R}(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s=j)\bar{R}_j(\boldsymbol{f}) = \sum_{j=1}^{K-|Y|} p(s=j)\bar{R}_j(\boldsymbol{f}) = R(\boldsymbol{f}).$$

Finally, let $\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y}) = \frac{1}{2^K-2}\sum_{Y\in\mathcal{Y}',Y\cap\bar{Y}=\emptyset}\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)$, we can obtain

$$\begin{aligned}
\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y}) &= \frac{1}{2^K-2}\sum_{Y\in\mathcal{Y}',Y\cap\bar{Y}=\emptyset}\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y) \\
&= \frac{1}{2^K-2}\sum_{Y\in\mathcal{Y}',Y\cap\bar{Y}=\emptyset}\left\{\sum_{y=1,y\in Y}^{K}\ell_y(\boldsymbol{x}) + \sum_{y=1,y\notin Y}^{K}\bar{\ell}_y(\boldsymbol{x})\right\} \\
&= \frac{1}{2^K-2}\left\{\sum_{y=1,y\notin\bar{Y}}^{K}\sum_{Y\in\mathcal{Y}',y\in Y}\ell_y(\boldsymbol{x}) + \sum_{y=1,y\notin\bar{Y}}^{K}\sum_{Y\in\mathcal{Y}',y\notin Y}\bar{\ell}_y(\boldsymbol{x}) + \sum_{y=1,y\in\bar{Y}}^{K}\sum_{Y\in\mathcal{Y}',y\notin Y}\bar{\ell}_y(\boldsymbol{x})\right\} \\
&= \frac{1}{2^K-2}\left\{\sum_{y=1,y\notin\bar{Y}}^{K}2^{K-j-1}\ell_y(\boldsymbol{x}) + \sum_{y=1,y\notin\bar{Y}}^{K}\left(2^{K-j-1}-1\right)\bar{\ell}_y(\boldsymbol{x}) + \sum_{y=1,y\in\bar{Y}}^{K}\left(2^{K-j}-1\right)\bar{\ell}_y(\boldsymbol{x})\right\} \\
&= \frac{2^{K-j-1}}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}\ell_y(\boldsymbol{x}) + \frac{2^{K-j-1}-1}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}\bar{\ell}_y(\boldsymbol{x}) + \frac{2^{K-j}-1}{2^K-2}\sum_{y=1,y\in\bar{Y}}^{K}\bar{\ell}_y(\boldsymbol{x}).
\end{aligned}$$

□

Appendix D
The Proof of Theorem 4

Based on the definitions of expected risk and empirical risk, i.e.,

$$\bar{R}(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s=j)\bar{R}_j(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s=j)\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\left[\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y})\right],$$

$$\bar{R}_n(\boldsymbol{f}) = \sum_{j=1}^{K-1} p(s=j)\bar{R}_j^n(\boldsymbol{f}) = \sum_{j=1}^{K-1} \frac{p(s=j)}{n_j}\sum_{i=1}^{n_j}\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}_i),\bar{Y}_i),$$

we can deduce the following lemma.

Lemma 5. *The inequality holds:* $R(\boldsymbol{f}_n) - R(\boldsymbol{f}^*) \leq 2\sum_{j=1}^{K-1} p(s=j)\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right|$.

Proof. Intuitively, we can obtain

$$\begin{aligned}
R(\boldsymbol{f}_n) - R(\boldsymbol{f}^*) &= \bar{R}(\boldsymbol{f}_n) - \bar{R}(\boldsymbol{f}^*) \\
&= \left[\bar{R}(\boldsymbol{f}_n) - \bar{R}_n(\boldsymbol{f}^*)\right] + \left[\bar{R}_n(\boldsymbol{f}_n) - \bar{R}_n(\boldsymbol{f}^*)\right] + \left[\bar{R}_n(\boldsymbol{f}^*) - \bar{R}(\boldsymbol{f}^*)\right] \\
&\leq \bar{R}(\boldsymbol{f}_n) - \bar{R}_n(\boldsymbol{f}_n) + \bar{R}_n(\boldsymbol{f}^*) - \bar{R}(\boldsymbol{f}^*) \quad \left(\because \bar{R}_n(\boldsymbol{f}_n) - \bar{R}_n(\boldsymbol{f}^*) \leq 0\right) \\
&= 2\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\bar{R}_n(\boldsymbol{f}) - \bar{R}(\boldsymbol{f})\right| \\
&= 2\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\sum_{j=1}^{K-1} p(s=j)\bar{R}_j^n(\boldsymbol{f}) - \sum_{j=1}^{K-1} p(s=j)\bar{R}_j(\boldsymbol{f})\right| \\
&\leq 2\sum_{j=1}^{K-1} p(s=j)\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right|.
\end{aligned}$$

□

Next, we will employ Rademacher Complexity to bound $\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right|$, as demonstrated below.

Lemma 6. *Let* $M_j = \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{f}\in\mathcal{F}}\bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y})$*, and* $\mathcal{H}_j = \{h : (\boldsymbol{x},\bar{Y}) \in (\mathcal{X}\times\bar{\mathcal{Y}}_j) \mapsto \bar{\mathcal{L}}_j(\boldsymbol{f}(\boldsymbol{x}),\bar{Y})|\boldsymbol{f}\in\mathcal{F}\}$ *is a class of measurable functions. For all* $j \in \{1,2,\ldots,K-1\}$ *and any* $\delta > 0$*, with a probability at least* $1-\delta$*,*

$$\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right| \leq 2\mathfrak{R}_{n_j}(\mathcal{H}_j) + \frac{M_j}{2}\sqrt{\frac{\log 2/\delta}{2n_j}},$$

where $\mathfrak{R}_{n_j}(\mathcal{H}_j) = \mathbb{E}_{\boldsymbol{x},\bar{Y},\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\sigma_i h(\boldsymbol{x}_i,\bar{Y}_i)\right]$ *represents the expected Rademacher Complexity of* $\mathcal{H}_j$*. Here,* $\boldsymbol{\sigma} = \{\sigma_1,\sigma_2,\ldots,\sigma_{n_j}\}$ *is a vector consisting of* n_j *Rademacher variables, taking values from* $\{-1,+1\}$ *with even probabilities.*

Proof. We first consider the single direction $\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right)$ with a probability at least $1-\frac{\delta}{2}$. Due to M_j being an upper bound for $\bar{\mathcal{L}}_j$, the change in $\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right)$ is no greater than $\frac{M_j}{2n_j}$ when we use an arbitrary instance $(\boldsymbol{x}_i', \bar{Y}_i')$ to replace an instance $(\boldsymbol{x}_i, \bar{Y}_i)$ belonging to $\bar{D}$. Based on McDiarmid's inequality [46], for any $\delta > 0$, with a probability at least $1-\frac{\delta}{2}$,

$$\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right) \leq \mathbb{E}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right)\right] + \frac{M_j}{2}\sqrt{\frac{\log 2/\delta}{2n_j}}.$$

By symmetrization [47], we have

$$\mathbb{E}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_j^n(\boldsymbol{f}) - \bar{R}_j(\boldsymbol{f})\right)\right] \leq 2\mathfrak{R}_{n_j}(\mathcal{H}_j).$$

□

Subsequently, we need to establish a bound of $\mathfrak{R}_{n_j}(\mathcal{H}_j)$ to get the bound of $R(\boldsymbol{f}_n) - R(\boldsymbol{f}^*)$.

Lemma 7. *Assuming $\ell_y(\boldsymbol{x})$ and $\bar{\ell}_y(\boldsymbol{x})$ are ρ^+-Lipschitz and ρ^--Lipschitz with respect to $\boldsymbol{f}(\boldsymbol{x})$ for any $y \in \mathcal{Y}$. Then, for all $j \in \{1, 2, \ldots, K-1\}$, we have*

$$\mathfrak{R}_{n_j}(\mathcal{H}_j) \leq \sqrt{2}K \left(\frac{2^{K-j-1}}{2^{K-2}} \rho^+ + \frac{3 \cdot 2^{K-j-1}}{2^{K-2}} \rho^- \right) \sum_{y=1}^{K} \mathfrak{R}_{n_j}\left(\mathcal{G}_y\right).$$

Proof. Initially, let $\ell \circ \mathcal{F}$ and $\bar{\ell} \circ \mathcal{F}$ refer to $\{\ell \circ \mathcal{F} | \boldsymbol{f} \in \mathcal{F}\}$ and $\{\bar{\ell} \circ \mathcal{F} | \boldsymbol{f} \in \mathcal{F}\}$, respectively. We can then employ the Rademacher vector contraction inequality [48] to derive the following results:

$$\begin{aligned}
\mathfrak{R}_{n_j}(\mathcal{H}_j) &= \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}_j} \frac{1}{n_j} \sum_{i=1}^{n_j} \sigma_i h(\boldsymbol{x}_i, \bar{Y}_i) \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \sigma_i \bar{\mathcal{L}}_j(\boldsymbol{x}_i, \bar{Y}_i) \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \sigma_i \left\{ \sum_{y=1, y \notin \bar{Y}_i}^{K} \frac{2^{K-j-1}}{2^K - 2} \ell_y(\boldsymbol{x}_i) + \sum_{y=1, y \notin \bar{Y}_i}^{K} \frac{2^{K-j-1} - 1}{2^K - 2} \bar{\ell}_y(\boldsymbol{x}_i) \right.\right. \\
&\left.\left. + \sum_{y=1, y \in \bar{Y}_i}^{K} \frac{2^{K-j} - 1}{2^K - 2} \bar{\ell}_y(\boldsymbol{x}_i) \right\} \right] \\
&\leq \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1}}{2^K - 2} \sigma_i \sum_{y=1, y \notin \bar{Y}_i}^{K} \ell_y(\boldsymbol{x}_i) \right] \\
&+ \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1} - 1}{2^K - 2} \sigma_i \sum_{y=1, y \notin \bar{Y}_i}^{K} \bar{\ell}_y(\boldsymbol{x}_i) \right] \\
&+ \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j} - 1}{2^K - 2} \sigma_i \sum_{y=1, y \in \bar{Y}_i}^{K} \bar{\ell}_y(\boldsymbol{x}_i) \right].
\end{aligned}$$

Here, we introduce random variables $\alpha_{i,y} = \mathbb{I}[y \notin \bar{Y}_i]$ for any $y \in \mathcal{Y}$, where $\mathbb{I}[\cdot]$ represents an indicator function. Then, we can express

$$\begin{aligned}
\mathfrak{R}_{n_j}(\mathcal{H}_j) &\leq \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1}}{2^K - 2} \sigma_i \sum_{y=1}^{K} \alpha_{i,y} \ell_y(\boldsymbol{x}_i) \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1} - 1}{2^K - 2} \sigma_i \sum_{y=1}^{K} \alpha_{i,y} \bar{\ell}_y(\boldsymbol{x}_i) \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j} - 1}{2^K - 2} \sigma_i \sum_{y=1}^{K} (1 - \alpha_{i,y}) \bar{\ell}_y(\boldsymbol{x}_i) \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1}}{2^K - 2} \sigma_i \sum_{y=1}^{K} \frac{1}{2} (2\alpha_{i,y} - 1 + 1) \ell_y(\boldsymbol{x}_i) \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1} - 1}{2^K - 2} \sigma_i \sum_{y=1}^{K} \frac{1}{2} (2\alpha_{i,y} - 1 + 1) \bar{\ell}_y(\boldsymbol{x}_i) \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{2^{K-j} - 1}{2^K - 2} \sigma_i \sum_{y=1}^{K} \frac{1}{2} (1 - 2\alpha_{i,y} + 1) \bar{\ell}_y(\boldsymbol{x}_i) \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1}}{2^K - 2} \left\{ \sum_{y=1}^{K} (2\alpha_{i,y} - 1) \sigma_i \ell_y(\boldsymbol{x}_i) + \sum_{y=1}^{K} \sigma_i \ell_y(\boldsymbol{x}_i) \right\} \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n_j} \sum_{i=1}^{n_j} \frac{2^{K-j-1} - 1}{2^K - 2} \left\{ \sum_{y=1}^{K} (2\alpha_{i,y} - 1) \sigma_i \bar{\ell}_y(\boldsymbol{x}_i) + \sum_{y=1}^{K} \sigma_i \bar{\ell}_y(\boldsymbol{x}_i) \right\} \right] + \\
&\mathbb{E}_{p(\boldsymbol{x}, \bar{Y} | s=j)} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n_j} \sum_{i=1}^{n_j} \frac{2^{K-j} - 1}{2^K - 2} \left\{ \sum_{y=1}^{K} (1 - 2\alpha_{i,y}) \sigma_i \bar{\ell}_y(\boldsymbol{x}_i) + \sum_{y=1}^{K} \sigma_i \bar{\ell}_y(\boldsymbol{x}_i) \right\} \right].
\end{aligned}$$

As $(1-2\alpha_{i,y})\sigma_i$ and $(2\alpha_{i,y}-1)\sigma_i$ have the same distribution as σ_i, we have

$$\begin{aligned}
\mathfrak{R}_{n_j}(\mathcal{H}_j) \leq\ & \mathbb{E}_{p(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{2^{K-j-1}}{2^K-2}\sum_{y=1}^{K}\sigma_i\ell_y(\boldsymbol{x}_i)\right]+\\
& \mathbb{E}_{p(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{2^{K-j-1}-1}{2^K-2}\sum_{y=1}^{K}\sigma_i\bar{\ell}_y(\boldsymbol{x}_i)\right]+\\
& \mathbb{E}_{p(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{2n_j}\sum_{i=1}^{n_j}\frac{2^{K-j}-1}{2^K-2}\sum_{y=1}^{K}\sigma_i\bar{\ell}_y(\boldsymbol{x}_i)\right]\\
=\ & \frac{2^{K-j-1}}{2^K-2}\mathbb{E}_{p(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\sum_{y=1}^{K}\sigma_i\ell_y(\boldsymbol{x}_i)\right]+\\
& \frac{3\cdot 2^{K-j-1}-2}{2^K-2}\mathbb{E}_{p(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\sum_{y=1}^{K}\sigma_i\bar{\ell}_y(\boldsymbol{x}_i)\right]\\
\leq\ & \frac{2^{K-j-1}}{2^K-2}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\sum_{y=1}^{K}\sigma_i\ell_y(\boldsymbol{x}_i)\right]+\\
& \frac{3\cdot 2^{K-j-1}-2}{2^K-2}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n_j}\sum_{i=1}^{n_j}\sum_{y=1}^{K}\sigma_i\bar{\ell}_y(\boldsymbol{x}_i)\right] \quad (\because p(\boldsymbol{x})=p(\boldsymbol{x}|s=j))\\
\leq\ & \frac{2^{K-j-1}}{2^K-2}\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\ell\circ\mathcal{F})+\frac{3\cdot 2^{K-j-1}-2}{2^K-2}\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\bar{\ell}\circ\mathcal{F})\\
=\ & \frac{2^{K-j-1}}{2^K-2}K\mathfrak{R}_{n_j}(\ell\circ\mathcal{F})+\frac{3\cdot 2^{K-j-1}-2}{2^K-2}K\mathfrak{R}_{n_j}(\bar{\ell}\circ\mathcal{F})\\
\leq\ & \sqrt{2}K\rho^{+}\frac{2^{K-j-1}}{2^K-2}\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\mathcal{G}_y)+\sqrt{2}K\rho^{-}\frac{3\cdot 2^{K-j-1}-2}{2^K-2}\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\mathcal{G}_y)\\
=\ & \sqrt{2}K\left(\frac{2^{K-j-1}}{2^K-2}\rho^{+}+\frac{3\cdot 2^{K-j-1}-2}{2^K-2}\rho^{-}\right)\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\mathcal{G}_y).
\end{aligned}$$

□

According to the above lemmas (Lemma 5, Lemma 6, and Lemma 7), we can establish the estimation error bound. For any $\delta>0$, with a probability at least $1-\delta$, we have

$$R(\boldsymbol{f}_n)-R(\boldsymbol{f}^{*})\leq\sum_{j=1}^{K-1}p(s=j)\left\{4\sqrt{2}K\left(\frac{2^{K-j-1}}{2^K-2}\rho^{+}+\frac{3\cdot 2^{K-j-1}-2}{2^K-2}\rho^{-}\right)\sum_{y=1}^{K}\mathfrak{R}_{n_j}(\mathcal{G}_y)+M_j\sqrt{\frac{\log 2/\delta}{2n_j}}\right\}.$$

Appendix E
The Details of Section 4

Here, we derive $\bar{\mathcal{L}}'(\boldsymbol{x},\bar{Y})$ by constructing an upper bound of $\bar{\mathcal{L}}_{\mathrm{MAE}}$, whose specific deducing process is shown in the below:

$$\begin{aligned}
\bar{\mathcal{L}}_{\mathrm{MAE}}(\boldsymbol{f}(\boldsymbol{x}),\bar{Y}) =\ & \frac{2^{K-|\bar{Y}|-1}}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}(1-f_y(\boldsymbol{x}))+\frac{2^{K-|\bar{Y}|-1}-1}{2^K-2}\sum_{y=1,y\notin\bar{Y}}^{K}f_y(\boldsymbol{x})+\frac{2^{K-|\bar{Y}|}-1}{2^K-2}\sum_{y=1,y\in\bar{Y}}^{K}f_y(\boldsymbol{x})\\
\leq\ & \frac{1}{2^K-2}\left\{\sum_{y=1,y\notin\bar{Y}}^{K}2^{K-|\bar{Y}|-1}(1-f_y(\boldsymbol{x}))+\sum_{y=1,y\notin\bar{Y}}^{K}2^{K-|\bar{Y}|-1}f_y(\boldsymbol{x})+\sum_{y=1,y\in\bar{Y}}^{K}2^{K-|\bar{Y}|}f_y(\boldsymbol{x})\right\}\\
\leq\ & \frac{1}{2^{K-1}}\left\{\sum_{y=1,y\notin\bar{Y}}^{K}2^{K-|\bar{Y}|-1}(1-f_y(\boldsymbol{x}))+\sum_{y=1,y\notin\bar{Y}}^{K}2^{K-|\bar{Y}|-1}f_y(\boldsymbol{x})+\sum_{y=1,y\in\bar{Y}}^{K}2^{K-|\bar{Y}|}f_y(\boldsymbol{x})\right\}\\
=\ & \sum_{y=1,y\notin\bar{Y}}^{K}2^{-|\bar{Y}|}(1-f_y(\boldsymbol{x}))+\sum_{y=1,y\notin\bar{Y}}^{K}2^{-|\bar{Y}|}f_y(\boldsymbol{x})+\sum_{y=1,y\in\bar{Y}}^{K}2^{1-|\bar{Y}|}f_y(\boldsymbol{x})
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{y=1, y \notin \bar{Y}}^{K} (1 - 2^{-|\bar{Y}|})(1 - f_y(\boldsymbol{x})) + \sum_{y=1, y \notin \bar{Y}}^{K} (1 - 2^{-|\bar{Y}|}) f_y(\boldsymbol{x}) + \sum_{y=1, y \in \bar{Y}}^{K} 2^{1-|\bar{Y}|} f_y(\boldsymbol{x}) \\
&\leq \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} (1 - f_y(\boldsymbol{x})) + \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} f_y(\boldsymbol{x}) + \sum_{y=1, y \in \bar{Y}}^{K} (e^{2^{1-|\bar{Y}|}} - 1) f_y(\boldsymbol{x}) \\
&\leq \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} (1 - f_y(\boldsymbol{x})) + \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} f_y(\boldsymbol{x}) + \sum_{y=1, y \in \bar{Y}}^{K} e^{2^{1-|\bar{Y}|}} f_y(\boldsymbol{x}) \\
&\leq - \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} \log(f_y(\boldsymbol{x})) - \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} \log(1 - f_y(\boldsymbol{x})) - \sum_{y=1, y \in \bar{Y}}^{K} e^{2^{1-|\bar{Y}|}} \log(1 - f_y(\boldsymbol{x})) \\
&= - \sum_{y=1, y \notin \bar{Y}}^{K} e^{2^{-|\bar{Y}|}} \{\log(f_y(\boldsymbol{x})) + \log(1 - f_y(\boldsymbol{x}))\} - \sum_{y=1, y \in \bar{Y}}^{K} e^{2^{1-|\bar{Y}|}} \log(1 - f_y(\boldsymbol{x})) \\
&= \bar{\mathcal{L}}'(\boldsymbol{x}, \bar{Y}).
\end{aligned}
$$

The sixth inequality above holds because of $1-z \leq e^{-z}$ and $1+z \leq e^{z}$. The eighth inequality can hold due to $1-z \leq -\log z$.